SVKM's

D. J. Sanghvi College of Engineering

Program: B.Tech in Computer Engineering **Academic Year: 2022** **Duration: 3 hours**
Date: 10.01.2023
Time: 10:30 am to 01:30 pm
Subject: Data Mining and Warehouse (Semester V) **Marks: 75**

Instructions: Candidates should read carefully the instructions printed on the question paper and on the cover page of the Answer Book, which is provided for their use.
(1) This question paper contains three pages.
(2) All Questions are Compulsory.
(3) All questions carry equal marks.
(4) Answer to each new question is to be started on a fresh page.
(5) Figures in the brackets on the right indicate full marks.
(6) Assume suitable data wherever required, but justify it.
(7) Draw the neat labelled diagrams, wherever necessary.

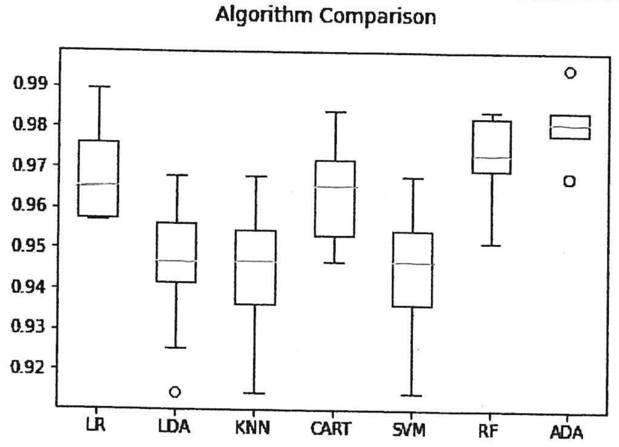
Question No.		**Max. Marks**
Q1 (a)	Define Data Warehouse. Explain the features of a data warehouse.	[10]
	OR	
Q1 (a)	Explain Data Mining as a step in KDD process. List applications of data mining	[10]
Q1 (b)	List major steps in ETL process	[05]
Q2 (a)	Compare Bagging and Boosting.	[07]
	OR	
Q2 (a)	Compare Partitioning Methods and Hierarchical Methods.	[07]
Q2 (b)	Design a Star schema for product sales considering dimensions like time, product, branch and location	[08]
Q3 (a)		

Figure above represents boxplots for accuracy values of 10-fold cross validations 7 listed algorithms on a dataset. [02]

i. Explain boxplot. [04]

ii. List the highest accuracy achieved by each model. [04]

iii. Discuss which model is best.

OR

Q3 (a)

i. Explain Discretization by Binning. [04]

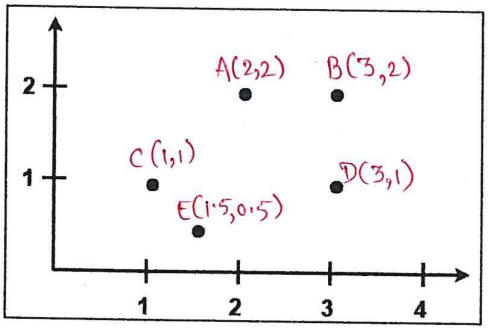
ii. For the data D= {4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34}

Perform following: No. of bins =3

1. Partition into equal-frequency (**equi-depth**) bins. [02]
2. Smoothing by **bin means**. [02]
3. Smoothing by **bin boundaries**. [02]

Q3 (b) Explain Market Basket Analysis [05]

Q4 (a) Identify and apply appropriate algorithm to Cluster the give data points into 2 clusters. [09]



OR

Q4 (a) Generate frequent Itemsets for following dataset using Apriori Algorithm and generate strong rules. Minimum support count = 2. Confidence = 60%. [09]

TID	items
T1	I1, I2 , I5
T2	I2,I4
T3	I2,I3
T4	I1,I2,I4
T5	I1,I3
T6	I2,I3
T7	I1,I3
T8	I1,I2,I3,I5
T9	I1,I2,I3

<table>
<tr><td>Q4 (b)</td><td>Define and Compute using the following confusion matrix:
i. Accuracy
ii. Precision
iii. Recall.
<table>
<tr><td>Actual
Predicted</td><td>Cancer = yes</td><td>Cancer = no</td></tr>
<tr><td>Cancer = yes</td><td>9</td><td>21</td></tr>
<tr><td>Cancer = no</td><td>14</td><td>956</td></tr>
</table>
</td><td>
[02]
[02]
[02]</td></tr>
<tr><td>Q5 (a)</td><td>Explain OLAP operations with example</td><td>[08]</td></tr>
<tr><td></td><td><b>OR</b></td><td></td></tr>
<tr><td>Q5 (a)</td><td>Discuss various scenarios where data warehouse is updated. Explain the process of application of these updates</td><td>[08]</td></tr>
<tr><td>Q5 (b)</td><td>Explain Web Content Mining</td><td>[07]</td></tr>
</table>

All the Best!